Analysis of House CVs

Manhui Xiao

Executive Summary

The dataset contained the house CV prices, house characteristics and SA1 statistics for 1051 houses in Auckland, as well as SA1 population information and deprivation index collected from the Koordinates API and the University of Otago respectively.

There are 17 variables in total, with Address being the address of the house and Suburb being a categorical variable denoting the suburb the house is in. Variables to “n – n years” represent the number of people within the age range living in the house’s SA1 area. The rest of the variables are all numerical and describe the house in more detail.

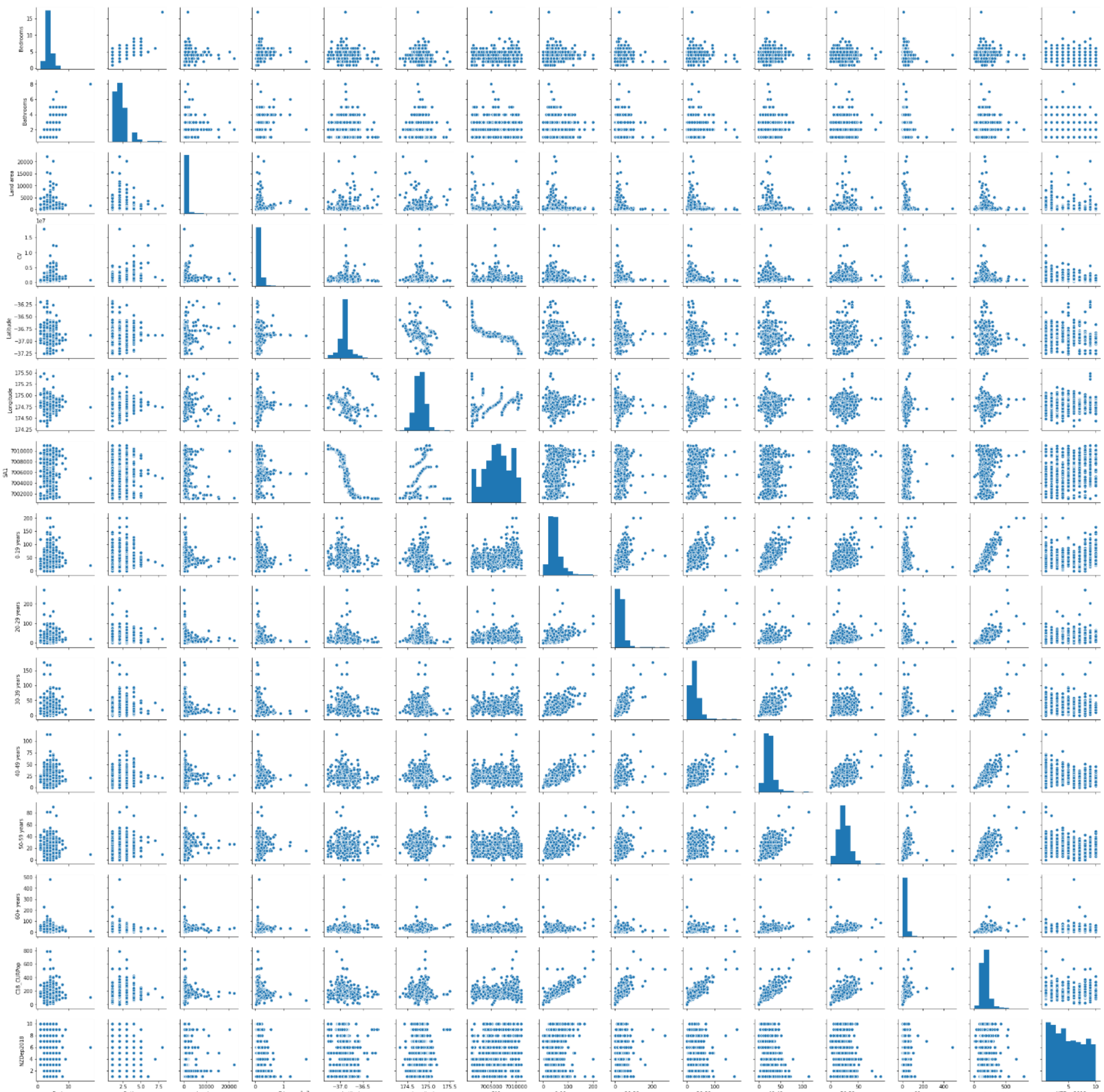
Through analytical exploration of the data as well as visualization of the correlation between variables, we attempt to find the model with the highest rate of accuracy for predicting the CV price of a house. The deprivation index and bathroom/bedroom count are found to play a major part in determining the price.

Initial data analysis

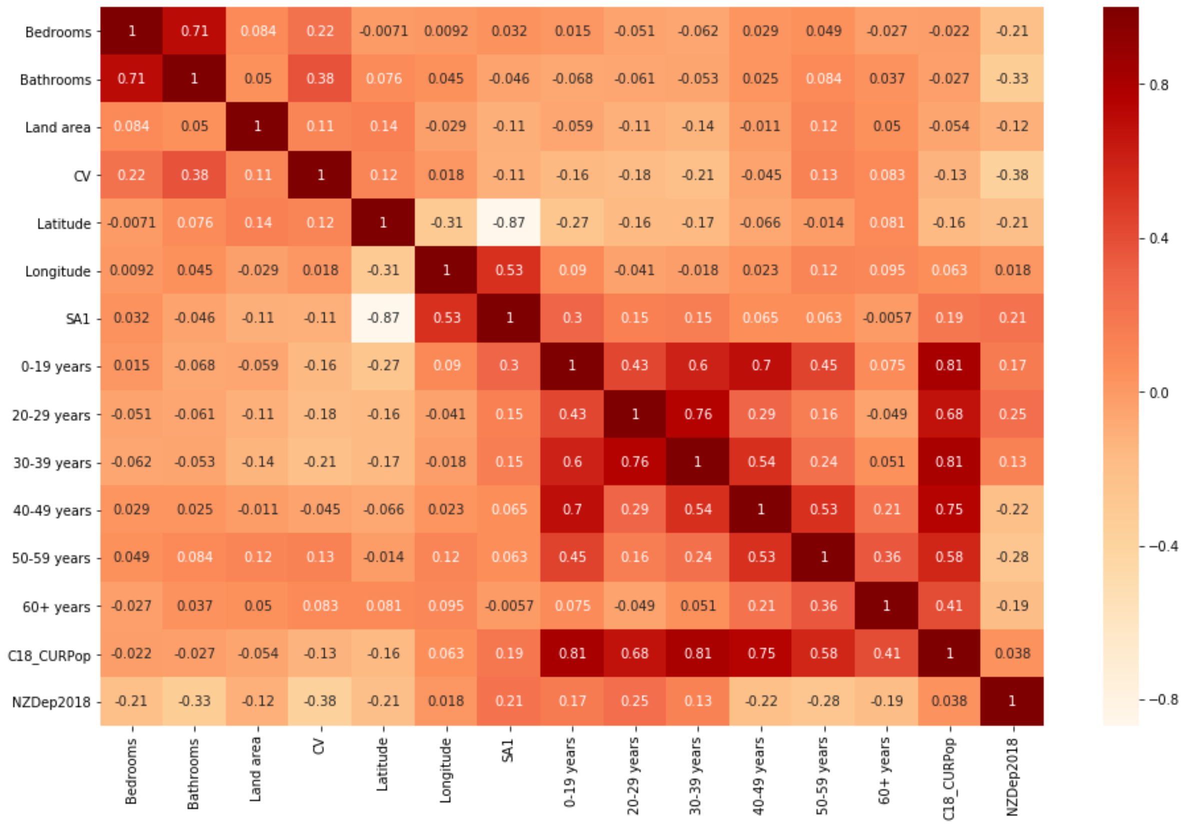
	Bedrooms	Bathrooms	Land area	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	C18_CURPop	NZDep2018
count	1051.000000	1049.000000	1051.000000	1.051000e+03	1051.000000	1051.000000	1.051000e+03	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000
mean	3.777355	2.073403	856.989534	1.387521e+06	-36.893715	174.799325	7.006319e+06	47.549001	28.963844	27.042816	24.125595	22.615604	29.360609	179.914367	5.063749
std	1.169412	0.992985	1588.156219	1.182939e+06	0.130100	0.119538	2.591262e+03	24.692205	21.037441	17.975408	10.942770	10.210578	21.805031	71.059280	2.913471
min	1.000000	1.000000	40.000000	2.700000e+05	-37.265021	174.317078	7.001130e+06	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	3.000000	1.000000
25%	3.000000	1.000000	321.000000	7.800000e+05	-36.950565	174.720779	7.004416e+06	33.000000	15.000000	15.000000	18.000000	15.000000	18.000000	138.000000	2.000000
50%	4.000000	2.000000	571.000000	1.080000e+06	-36.893132	174.798575	7.006325e+06	45.000000	24.000000	24.000000	24.000000	21.000000	27.000000	174.000000	5.000000
75%	4.000000	3.000000	825.000000	1.600000e+06	-36.855789	174.880944	7.008384e+06	57.000000	36.000000	33.000000	30.000000	27.000000	36.000000	210.000000	8.000000
max	17.000000	8.000000	22240.000000	1.800000e+07	-36.177655	175.492424	7.011028e+06	201.000000	270.000000	177.000000	114.000000	90.000000	483.000000	789.000000	10.000000

Initial descriptive statistics were generated for each of the variables and shown above. Non-numerical variables were excluded from the descriptive statistics as it is impossible to calculate mean, std etc for a categorical variable.

Analysis of correlations and patterns in the data



Most of the data seems to follow the same general patter for their matching group. Due to the many variables that are available, it is difficult to determine whether or not a point is an outlier.



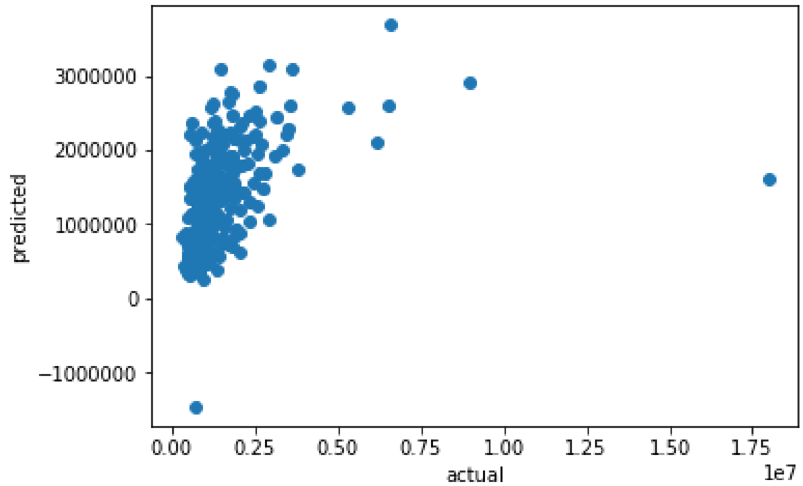
The correlations between the numerical columns are shown in the above heatmap. We can see a strong positive correlation between C18_CURPop and the n-n years variables, which is not surprising as they all refer to population-related statistics. The same can be said for Longitude/Latitude and their strong correlation to SA1, since they all relate to location. There does not seem to be much correlation to CV however, other than the moderately strong relations of Bedroom/Bathroom and deprivation.

Build a model and comment on it

Multiple linear regressions will be made on this dataset with an attempt to optimize the data to obtain higher accuracy rates. There were 3 rows with null values in the dataset:

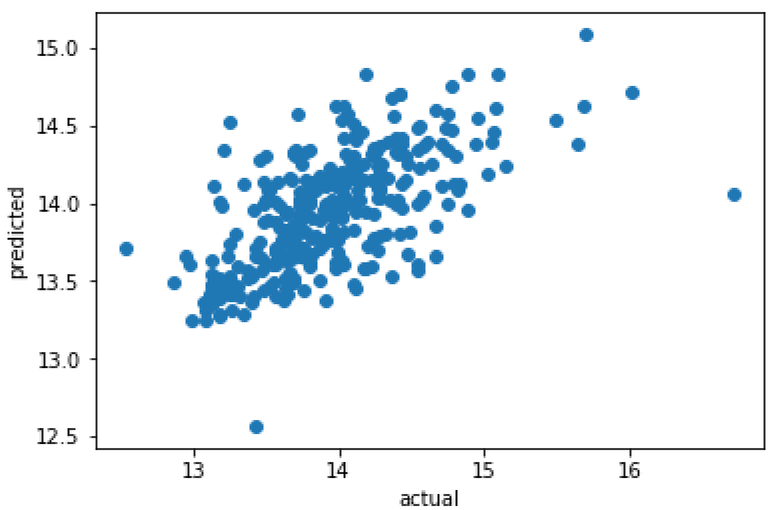
	Bedrooms	Bathrooms	Address	Land area	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	Suburbs	C18_CURPop	NZDep2018
309	4	NaN	14 Hea Road Hobsonville, Auckland	214.0	1250000	-36.798371	174.647430	7002267	60	66	60	24	24	18	Hobsonville	252	2.0
311	4	NaN	16 Hea Road Hobsonville, Auckland	245.0	1100000	-36.798371	174.647430	7002267	60	66	60	24	24	18	Hobsonville	252	2.0
568	1	1.0	14 Te Rangitawhiri Road Great Barrier Island, ...	2141.0	740000	-36.197282	175.416921	7001131	27	6	6	18	39	60	NaN	156	9.0

For the initial model I removed the 2 rows with bathroom as the null value but kept the other one as its null value is in the Suburbs column which is categorical and not included in the model anyways. The model yielded a r^2 score of 0.161 and a mean squared error of 1496258333235.71 which is very inaccurate – especially as most of the prices aren't this high. A plot of predicted vs actual test data can be seen below. It can be seen that one price was somehow predicted as negative due to how low it was.



Model 2:

Since prices are often skewed, I transformed it with a log function, giving a correlation map with stronger correlations than the previous one. The model built with this data gave a r^2 score of 0.382, and mean squared error of 0.188 – considerably better than the previous model.



Model 3:

In the previous models, none of the string variables were included, however categorical variables are also likely to affect the model. The address variable is unlikely to be useful as it basically functions as an "ID" and there is only one of each. However, the suburb category can be used by encoding them to 0s and 1s. For this purpose I also removed the last null row from the database. However, possibly due to the amount of suburbs that were included, the model produced was not as accurate as model 2 with a r^2 of 0.181 and mean squared error of 0.248.

	Bedrooms	Bathrooms	Address	Land area	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	...	Suburbs_Waterview	Suburbs_Wattle Downs	Suburbs_Wellsford	Suburbs_Wesley	Suburbs_West Harbour	Suburbs_Westmere	Suburbs_Weymouth	Suburbs_Whe
0	5	3.0	106 Lawrence Crescent Hill Park, Auckland	714.0	960000	-37.012920	174.904069	7009770	48	27	...	0	0	0	0	0	0	0	
1	5	3.0	8 Corsica Way Karaka, Auckland	564.0	1250000	-37.063672	174.922912	7009991	42	18	...	0	0	0	0	0	0	0	
2	6	4.0	243 Harbourside Drive Karaka, Auckland	626.0	1250000	-37.063580	174.924044	7009991	42	18	...	0	0	0	0	0	0	0	
3	2	1.0	2/30 Hardington Street Onehunga, Auckland	65.0	740000	-36.912996	174.787425	7007871	42	6	...	0	0	0	0	0	0	0	
4	3	1.0	59 Israel Avenue Clover Park, Auckland	601.0	630000	-36.979037	174.892612	7008902	93	27	...	0	0	0	0	0	0	0	

Model 4:

For the final model I explored the addition and removal of various variables to make the prediction result more accurate. In the end I settled with the variables Bedrooms, Bathrooms, Land area, NZDep2018, SA1 and C_18CURPop, producing a model with a r^2 score of 0.371 and mean squared error of 0.191. Which this model did not improve the accuracy of tested results, these variables were those that had the most importance in the correlation map and the accuracy also did not decrease by much after removing the others – meaning they did not affect the output much. The biggest effect on accuracy was when I removed NZDep2018.

Conclusions

It is difficult to make an accurate prediction of the CV of houses based on the given variables using linear regression as not many of the variables have a strong correlation with the houses' prices. The suggested model to use if one was going to make a prediction would be the final model, using prices transformed using a log function.